A distribution-free two-sample run test applicable to high dimensional data

Article *in* Biometrika · November 2014

DOI: 10.1093/biomet/asu045

3 authors:

Munmun Biswas
7 PUBLICATIONS 199 CITATIONS
SEE PROFILE

Minerva Mukhopadhyay
Indian Institute of Technology Kanpur
16 PUBLICATIONS 203 CITATIONS
SEE PROFILE

Anil K. Ghosh
Indian Statistical Institute
66 PUBLICATIONS 1,125 CITATIONS
SEE PROFILE

A DISTRIBUTION-FREE TWO-SAMPLE RUN TEST APPLICABLE TO HIGH DIMENSIONAL SMALL SAMPLE SIZE DATA

Munmun Biswas^{+}, Minerva Mukhopadhyay$^{\circ}$ and Anil K. Ghosh^{+}

$^{+}$Theoretical Statistics and Mathematics Unit $^{\circ}$Applied Statistics Unit
Indian Statistical Institute, 203, B. T. Road, Kolkata 700108, India.

E-mail : munmun.biswas08@gmail.com, minervamukherjee@gmail.com, akghosh@isical.ac.in.

Abstract

Several multivariate generalizations of univariate nonparametric two-sample tests have been proposed in the literature. But, most of these generalizations fail to retain the distribution-free property of the univariate tests in the general multivariate set up. In this article, using the idea of shortest Hamiltonian path, we propose a multivariate generalization of the univariate run test. Unlike Friedman and Rafsky's (1979) multivariate run test, this proposed test has the distribution-free property in finite sample situations. Most of the existing two-sample tests perform poorly for high dimensional data, especially when the training sample size is small, and many of them are not applicable when the dimension of the data exceeds the sample size. But our proposed test can be conveniently used in high dimension low sample size situations. We investigate the power properties of the proposed test when the sample size remains fixed and dimension of the data grows to infinity. Several simulated and real data sets are also analyzed to evaluate its performance. Using our theoretical and empirical studies, we demonstrate the superiority of this proposed test over Friedman and Rafsky's run test and several other two-sample tests available in the literature.

Keywords: Distribution-free property, Level and power of a test, Shortest Hamiltonian path, Two-sample run test, Weak law of large numbers.

1 Introduction

In a two-sample testing problem, we test the equality of two distributions F and G based on two sets of independent observations $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m \overset{i.i.d.}{\sim} F$ and $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n \overset{i.i.d.}{\sim} G$. It is a well studied problem, and several nonparametric methods have been proposed for it. The Wilcoxon-Mann-Whiteny rank test, the Kolmogorov-Smirnov maximum deviation test and the Wald-Wolfowitz run test (see e.g., Gibbons and Chakraborti, 2003) are some examples of classical univariate rank based tests available in the literature. While the first one is mainly used when the alternative hypothesis suggests a stochastic ordering between F and G, the other two are used for more general alternatives. One useful feature of these tests is their distribution-free property.

Several attempts have been made in the literature to generalize these nonparametric tests to multivariate set up. Mardia (1967) developed a distribution-free two-sample location test for bivariate data. Multivariate rank based tests for location problem include Puri and Sen (1971), Randles and Peters (1990), Hetmansperger and Oja (1994), Möttönen and Oja (1995), Choi and Marden (1997) and Hettmansperger *et al.* (1998). A summary of these methods can be obtained in Oja and Randles (2004) and Oja (2010). However, unlike the univariate methods, these tests do not have the distribution-free property, and none of them can be used when the sample size is smaller than the dimension of the data. Liu and Singh (1993) used simplicial depth to develop two separate distribution-free tests, one for two-sample location problem and the other for two-sample scale problem. Depth based distribution-free tests for location and scale models were also considered in Rousson (2002), where he used Mahalanobis depth. But, these depth based tests cannot be used when the dimension of the data exceeds the sample size.

Multivariate nonparametric methods have been developed for general two-sample problems as well, but most them do not have the distribution-free property. Even the most natural generalizations of the Kolmogorov-Smirnov test is not distribution-free in two or higher dimensions (see e.g., Bickel, 1969). Friedman and Rafsky (1979) proposed another multivariate generalization of the Kolmorov-Smirnov statistic based on the idea of minimum spanning tree (MST). Using the concept of MST, they also generalize the Wald-Wolfowtz run test to multivariate setup. Schilling (1986) and Henze (1988) proposed multivariate two-sample test based on nearest neighbor type coincidences. Other multivariate nonparametric tests for general two-sample problem include Hall and Tajvidi (2002), Zech and Aslan (2003), Baringhaus and Franz (2004, 2010) and Liu and Modarres (2011). These tests can be used in high dimension small sample size situations, but none of them are distribution-free. In these cases, one either uses the conditional test based on the permutation principle or the large sample test based on the asymptotic null distribution of the test statistic. Ferger (2000) proposed a distribution-free two-sample test from the perspective of change point detection. However, for proper implementation of this test, one needs to find a suitable weight function and an appropriate asymmetric kernel function. Rosenbaum (2005) proposed a simple exact sample distribution-free test using the idea of an optimal non-bipartite matching (see e.g., Lu *et. al.*, 2011). This test can also be used for high dimensional data if the Euclidean metric is used for computing the pairwise distances between the sample observations.

In this article, we propose a multivariate generalization of the Wald-Wolfowitz run test using the idea of shortest Hamiltonian path (SHP). Like Friedman and Rafsky's run test (henceforth, we will refer to it as the FR test), this test is also based on pairwise distances between the observations, and it can be conveniently used for high dimensional low sample size data or even for functional data taking values in a Banach space. Compared to the FR test, it enjoys better power properties in the high dimension low sample size set up, which we will see later. Also, unlike the FR test, this proposed method has the finite sample distribution-free property. In fact, the null distribution of our test statistic exactly matches with that of the univariate run statistic. So, in that sense, it can be viewed as the most natural generalization of the univariate run test. The description of this test is given in the next section.

2 Multivariate run test based on shortest Hamiltonian path

Consider a graph $\mathcal{G}$ on N vertices. A Hamiltonian path in $\mathcal{G}$ is defined as a connected, acyclic sub-graph of $\mathcal{G}$ with $N-1$ edges, where no vertex has degree bigger than two. In other words, a Hamiltonian path is a path that visits each vertex of $\mathcal{G}$ exactly once. For any given $\mathcal{G}$, a Hamiltonian path may or may not exist, but if $\mathcal{G}$ is a complete graph on N vertices, there are $N!$ Hamiltonian paths. However, for every path, there is another path in the reverse order. So, if we consider them as the same path, there are $N!/2$ distinct Hamiltonian paths. Now, consider $\mathcal{G}$ to be a complete graph on N vertices, where each of the $\binom{N}{2}$ edges has a cost (e.g., the distance between the two vertices of the edge) associated with it. Now, for each Hamiltonian path, one can compute the sum of the costs corresponding to its $N-1$ edges, which is defined to be the cost of the Hamiltonian path. The Hamiltonian path having the minimum cost is defined as the shortest Hamiltonian path (SHP). For a graph $\mathcal{G}$, SHP may not be unique, but if the costs corresponding to different edges come from continuous distributions, it becomes unique with probability one. Figure 1 shows a complete graph on four vertices $\{\mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3, \mathbf{z}_4\}$ along with the costs corresponding to different edges. There are 12 distinct Hamiltonian path in this graph, but the path $\mathbf{z}_2 \to \mathbf{z}_1 \to \mathbf{z}_3 \to \mathbf{z}_4$ (or $\mathbf{z}_4 \to \mathbf{z}_3 \to \mathbf{z}_1 \to \mathbf{z}_2$) has the minimum cost. So, it is the SHP in this graph.

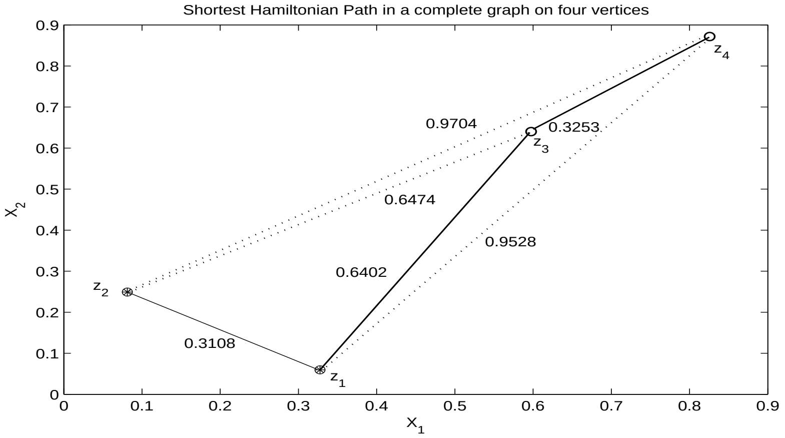

Figure 1: Shortest Hamiltonian path in a complete graph on four vertices

Suppose that we have m independent observations $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m$ from a d-dimensional distribution F and n independent observations $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n$ from another d-dimensional distribution G. Define $\mathbf{z}_i = \mathbf{x}_i$ for $i = 1, 2, \ldots, m$, and $\mathbf{z}_{m+i} = \mathbf{y}_i$ for $i = 1, 2, \ldots, n$. Now, consider a complete graph with $N = m + n$ vertices $\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_N$, where the edge between $\mathbf{z}_i$ and $\mathbf{z}_j$ $(1 \le i < j \le N)$ has the cost $\|\mathbf{z}_i - \mathbf{z}_j\|$, the Euclidean distance between $\mathbf{z}_i$ and $\mathbf{z}_j$. Let $\mathcal{H}$ be the SHP of this graph. We count the number of runs along $\mathcal{H}$, where a sequence of consecutive observations from the same distribution is considered as a run. In Figure 1, if we consider $\mathbf{z}_1, \mathbf{z}_2$ are from F and $\mathbf{z}_3, \mathbf{z}_4$ are from G, the number of runs turns out to be 2 ($\mathbf{z}_2 \to \mathbf{z}_1$ and $\mathbf{z}_3 \to \mathbf{z}_4$). Note that if F and G are widely separated, one would expect this number of runs to be small, while under H_0 it is expected to be large. So, we compute the number of runs $T_{m,n} = 1 + \sum_{i}^{N-1} U_i$, where U_i is an indicator variable that takes the value 1 if and only if the i-th edge of $\mathcal{H}$ connects two observations from two different distributions, and reject H_0 for small values of $T_{m,n}$. Note that if F and G are one-dimensional distributions, $\mathcal{H}$ is obtained by joining the observations $\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_N$ in increasing/decreasing order, and in that case, $T_{m,n}$ matches with the univariate run statistic. Therefore, just like the FR test, our proposed test can be viewed as another multivariate generalization of the univariate run test.

2.1 Exact and large sample distributions of $T_{m,n}$

From the above discussion, it is quite clear that the proposed test has the distribution-free property in one dimension, where it matches with the univariate run test and the FR test. But, unlike the FR test, our proposed generalization successfully retains this distribution-free property of the univariate run test in higher dimensions. Note that $T_{m,n}$ and the univariate run statistic are the same function of the ranks of the observations; the only difference is that in $T_{m,n}$, the ranks are computed along $\mathcal{H}$. Now, under H_0, because of the exchangeability of $\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_N$, irrespective of the underlying distribution and data dimension, this rank vector has the same distribution as in the univariate case. Therefore, $T_{m,n}$ has the distribution-free property, and its null distribution exactly matches with that of the univariate run statistic, which is given by

$$\begin{array}{lll} P(T_{m,n} = 2k) & = 2\binom{m-1}{k-1}\binom{n-1}{k-1}/\binom{N}{m} & \text{for } \; k = 1, 2, \ldots, \min\{m, n\} \text{ and} \\ P(T_{m,n} = 2k-1) & = \left[\binom{m-1}{k-1}\binom{n-1}{k-2} + \binom{m-1}{k-2}\binom{n-1}{k-1}\right]/\binom{N}{m} & \text{for } \; k = 2, \ldots, \min\{m, n\} + 1, \end{array}$$

where $\binom{a}{b} = 0$ if $a < b$ (see e.g., Wald and Wolfowitz, 1940; Gibbons and Chakraborti, 2003). In that sense, $T_{m,n}$ can be viewed as the most natural multivariate generalization of the univariate run statistic. If m and n are small, in order to carry out our test, we can use the statistical tables available for the univariate run test. Because of the discrete nature of $T_{m,n}$, one may need to use randomized tests to match the size of the test with the level of significance. However, if m and n are large, it may be computationally difficult to use the test based on the exact distribution of $T_{m,n}$. In that case, we can use the test based on the asymptotic null distribution of $T_{m,n}$. Again, this large sample distribution is the same as the large sample distribution of the univariate run statistic. Under H_0, the expectation and the variance of $T_{m,n}$ are given by $E_{H_0}(T_{m,n}) = \frac{2mn}{N} + 1$ and $V_{H_0}(T_{m,n}) = \frac{2mn(2mn-N)}{N^2(N-1)}$, respectively. Let us assume that as $N \to \infty$, $m/N \to \lambda$ for some $\lambda \in (0, 1)$. Under this condition, $E_{H_0}(T_{m,n}/N) \to 2\lambda(1-\lambda)$ and $V_{H_0}(T_{m,n}/\sqrt{N}) \to 4\lambda^2(1-\lambda)^2$ as $N \to \infty$. In such cases, one can show that (see e.g., Wald and Wolfowitz, 1940), under H_0,

$$T^*_{m,n} = \sqrt{N}\left[\frac{T_{m,n}/N - 2\lambda(1-\lambda)}{2\lambda(1-\lambda)}\right] \to N(0, 1).$$

2.2 Computation of $T_{m,n}$ for multivariate data

Unless m and n are very small, finding the SHP in a complete graph $\mathcal{G}$ is a computationally hard problem. In fact, it is equivalent to the well-known travelling salesman's problem, which is an NP-complete problem (see e.g., Garey, 1979). However, there are some good heuristic search algorithms available in literature (see e.g., Lawler et. al., 1985). In this article, we have adopted a popular method based on Kruskal's algorithm (see e.g., Kruskal, 1956). First, it sorts the edges of $\mathcal{G}$ in increasing order of their costs. Next, it starts from the edge with the minimum cost, and selects the edges one by one according to their costs. However, if an edge along with the previously chosen edges makes a cycle or if it makes the degree of a vertex more than two, we do not select that edge. The algorithm terminates when $N-1$ edges are chosen. The Hamiltonian path formed by these $N-1$ edges are considered as the SHP. We also tried two other heuristic algorithms from Lawler et. al. (1985) and the integer programming algorithm for finding SHP, but they did not make any visible difference in the performance of our test.

3 An illustrative example with high dimensional data

We have already seen that our proposed multivariate run test (henceforth referred to as the MR test) has an advantage over the FR test as far as the distribution-free property is concerned. Note that both the MR test and the FR test are based on pairwise distances between the sample observations. So, they are invariant under location change, rotation and homogeneous scale transformation. Also, these tests can be used for very high dimensional data even when the data dimension exceeds the sample size. Now, we consider a simple example to investigate how these tests perform in high dimension low sample size situations. Let us consider a two-sample problem, where the observations in F and G are distributed as $N_d((\mu,\ldots,\mu)^{'},\sigma^2\mathbf{I}_d)$ and $N_d((0,\ldots,0)^{'},\mathbf{I}_d)$, respectively. Here, N_d stands for a d-variate normal distribution, and $\mathbf{I}_d$ denotes the $d \times d$ identity matrix. Here, we consider two different choices of μ and σ^2, namely, ($\mu = 0.3$, $\sigma^2 = 1$) and ($\mu = 0$, $\sigma^2 = 1.3$), which lead to a location problem and a scale problem, respectively. In each case, we generated 20 observations from each distribution to test the null hypothesis $H_0 : F = G$ against the alternative $H_1 : F \neq G$. Each experiment was repeated 500 times, and the proportion of times a test rejected H_0 was considered as an estimate of its power. These estimated powers were computed for two multivariate run tests, the MR test and the FR test. Note that the FR test does not have the distribution-free property. So, in the case of FR test, we used the conditional test based on the permutation principle, where 500 permutations were used to estimate the cut off. We considered different values of d ranging from 3 to 3000, and the results are presented in Figure 2. Like the MR test, Rosenbaum's (2005) test based on non-bipartite matching (NBM) is applicable to general two-sample problem, and it also enjoys the distribution-free property. We have also considered this test for comparison. In future, we will refer to it as the NBM test. Note that this test needs to compute Mahalanobis distances between the observations, which involve inversion of the estimated pooled dispersion matrix. But this inverse does not exist if the dimension of the data exceeds the sample size. Therefore, to make it applicable to high dimensional data, instead of Mahalanobis distance, we used the Euclidean distance. Figure 2 also shows the estimated powers of this test. Here, all these tests have the 5% nominal level.

Note that both in location and scale problems, as d increases, the separability between F and G also increases. So, one should expect the powers of these tests to tend to unity as the

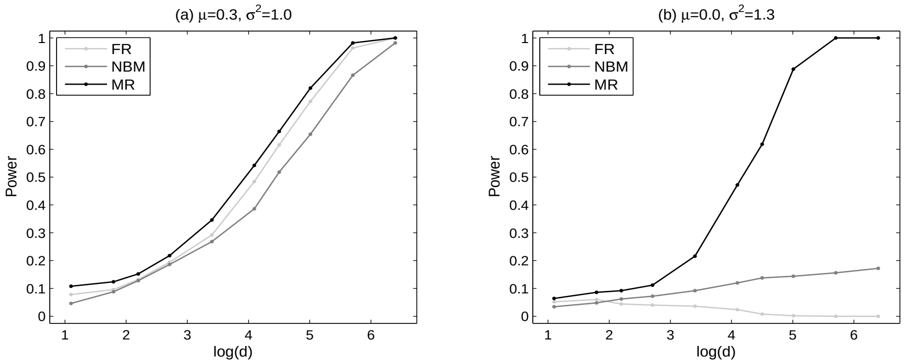

Figure 2: Powers of FR, NBM and MR tests for varying choices of d.

dimension increases. We observed that in the location problem (see Figure 2(a)), but not in the scale problem (see Figure 1(b)). In the location problem, the FR test and the MR test had comparable performance, though the latter one had an edge. Both of these run tests had higher powers than the NBM test. But, the result was more interesting in the case of scale problem. In this case, the MR test had substantially better performance than the other two tests. The powers of the MR test and the NBM test both increased with d, but the latter one increased at a very slow rate. The behavior of the FR test was more surprising. While the power of the MR test increased to unity, that of the FR test dropped down to zero as dimension increased. In the next section, we investigate the reasons behind such diametrically opposite behavior of these two multivariate run tests in high dimensional data.

4 Behavior of the multivariate run tests for high dimensional data

Here, we carry out a theoretical investigation to study the behavior of the MR test and the FR test for high dimensional data. In order to carry out this investigation, here we assume the sample sizes m and n to be fixed, and study the limiting behavior of the power functions of these tests as the dimension d diverges to infinity. In usual large sample asymptotics, we assume the dimension of the data to be fixed, and we expect to get more information about the separability between F and G as the sample sizes increase, but here we consider the sample sizes to be fixed, and we expect to get more information as the dimension increases. Here, we assume that we have m independent observations on $\mathbf{X} = (X^{(1)}, X^{(2)}, \ldots, X^{(d)})'$ from F, n independent observations on $\mathbf{Y} = (Y^{(1)}, Y^{(2)}, \ldots, Y^{(d)})'$ from G, and the observations on $\mathbf{X}$ and $\mathbf{Y}$ are also independent. Following Hall, Marron and Neeman (2005), we also make the following assumptions

(A1) *Fourth moments of $X^{(q)}$ and $Y^{(q)}$ are uniformly bounded.*

(A2) *Let $\mathbf{X}_1, \mathbf{X}_2$ be two independent copies of $\mathbf{X}$, and $\mathbf{Y}_1, \mathbf{Y}_2$ be two independent copied of $\mathbf{Y}$. Under some permutation of the $X^{(q)}$s (and the same permutation of the $Y^{(q)}$s), for $(U^{(q)}, V^{(q)}) = (X_1^{(q)}, X_2^{(q)}), (X_1^{(q)}, Y_1^{(q)})$ and $(Y_1^{(q)}, Y_2^{(q)})$, the sequence $\{(U^{(q)} - V^{(q)})^2, q \geq 1\}$ is ρ mixing, i.e., $\sup_{|q-q'|>r} |corr\{(U^{(q)} - V^{(q)})^2, (U^{(q')} - V^{(q')})^2\}| \leq \rho(r)$ where $\rho(r) \to 0$ as $r \to \infty$.*

(A3) *There exist constants* $\sigma_1^2, \sigma_2^2 > 0$ *and* ν^2 *such that (i)* $d^{-1}\sum_{q=1}^{d} Var(X^{(q)}) \to \sigma_1^2$, *(ii)* $d^{-1}\sum_{q=1}^{d} Var(Y^{(q)}) \to \sigma_2^2$ *and (iii)* $d^{-1}\sum_{q=1}^{d}\left[E(X^{(q)}) - E(Y^{(q)})\right]^2 \to \nu^2$ *as* $d \to \infty$.

Hall et. al. (2005) looked at the d-dimensional observations as infinite time series truncated at time d and studied the behavior of the inter-point distances as d increases. However, here we look at these observations from a multivariate data perspective. So, here we make some minor modifications to the assumptions of Hall et. al. (2005) (particularly, in (A2)). Ahn *et al.* (2007) also studied the geometry of high dimension low sample size data under slightly mild conditions. Andrews (1988) and de Jong (1995) assumed similar conditions to derive the weak law of large numbers (WLLN) for mixingales. Jung and Marron (2009) also assumed similar conditions for the large dimensional consistency of estimated principal component directions. Under the assumptions on uniformly bounded moments (A1) and weak dependence among component variables (A2), we have the WLLN for the sequence $\{(U^{(q)} - V^{(q)})^2,\ q \geq 1\}$ (the proof is straight-forward, and hence it is omitted). Again, depending on the choice of $(U^{(q)}, V^{(q)})$ $\left[(U^{(q)}, V^{(q)}) = (X_1^{(q)}, X_2^{(q)}), (X_1^{(q)}, Y_1^{(q)})\right.$ or $\left.(Y_1^{(q)}, Y_2^{(q)})\right]$, under the assumption (A3), $d^{-1}\sum_{q\geq 1} E(U^{(q)} - V^{(q)})^2$ converges to $2\sigma_1^2$, $2\sigma_2^2$ or $\sigma_1^2 + \sigma_2^2 + \nu^2$. So, under (A1)-(A3), as d tends to infinity, the pairwise distance between any two independent observations, when divided by $d^{1/2}$, converges in probability to positive constant as d tends to infinity. If both of them are from the same distribution, it converges to $\sigma_1\sqrt{2}$ or $\sigma_2\sqrt{2}$ depending on whether they are from F or G. If one of them is from F and the other one is from G, it converges to $\sqrt{\sigma_1^2 + \sigma_2^2 + \nu^2}$. Note that if the components of $\mathbf{X}$ and $\mathbf{Y}$ vectors are independent and identically distributed, as it was the case in the examples with normal distributions in Section 3, (A2) and (A3) hold automatically. Further, instead of (A1), we only need the finiteness of second order moments of the component variables for the convergence of these three types of inter-point distances. Under the assumptions (A1)-(A3), if $\nu^2 > 0$ or $\sigma_1^2 \neq \sigma_2^2$, the power of the MR test converges to unity as the dimension increases. This result is stated below as Theorem 1.

Theorem:*Assume that* F *and* G *both satisfy the assumptions (A1)-(A3). Also assume that* m *and* n *are such that* $m!\ n!/(m+n-1)! \leq \alpha$. *If* $\nu^2 > 0$ *and/or* $\sigma_1^2 \neq \sigma_2^2$, *the power of the MR test of level* α *converges to* 1 *as* $d \to \infty$.

Proof Recall that for any fixed m and n, $T_{m,n}$ has the same null distribution as the univariate run statistic, and hence one can check that $P_{H_0}(T_{m,n} \leq 3) = m!\ n!/(m+n-1)!$. Since $P_{H_0}(T_{m,n} \leq 3) \leq \alpha$, both $T_{m,n} = 2$ and $T_{m,n} = 3$ lead to the rejection of H_0. So, it is enough to prove that under the alternative H_1, $P_{H_1}(T_{m,n} > 3) \to 0$ as $d \to \infty$.

Now, note that as $d \to \infty$, $\|\mathbf{x}_i - \mathbf{x}_j\|/\sqrt{d} \xrightarrow{P} \sigma_1\sqrt{2}$ for $1 \leq i < j \leq m$, $\|\mathbf{y}_i - \mathbf{y}_j\|/\sqrt{d} \xrightarrow{P} \sigma_2\sqrt{2}$ for $1 \leq i < j \leq n$, and $\|\mathbf{x}_i - \mathbf{y}_j\|/\sqrt{d} \xrightarrow{P} \sqrt{\sigma_1^2 + \sigma_2^2 + \mu^2}$ for $1 \leq i \leq m$ and $1 \leq j \leq n$. For the ease of notation, here we denote these three limiting distances $\sigma_1\sqrt{2}, \sigma_2\sqrt{2}$ and $\sqrt{\sigma_1^2 + \sigma_2^2 + \mu^2}$ by a, b and c respectively. Note that $2c \geq a + b$, where the equality holds if and only if $\nu^2 = 0$ and $\sigma_1^2 = \sigma_2^2$. Let $\mathcal{H}$ be the SHP in the graph on $m + n$ vertices. Now, $\mathcal{H}$ can either (i) start and end with observations from same distribution or (ii) start with an observation from one distribution and end with an observation from the other distribution. Let us consider these two cases separately.

In case (i), $T_{m,n}$ can take only odd values, i.e., $T_{m,n} = 2k+1$ for some positive integer k. Now, if $\mathcal{H}$ starts and ends with observations from F. one can check that $\mathcal{H}$ contains $m-k$ '**XX**-type' edges, $n-k-1$ '**YY**-type' edges and $2k$ '**XY**-type' edges. So, the total cost of $\mathcal{H}$ converges (in probability) to $(m-k)a+(n-k-1)b+2kc = ma+(n-1)b+k(2c-a-b)$. Similarly, if $\mathcal{H}$ starts and ends with observations from G, the total cost of $\mathcal{H}$ converges to $(m-k-1)a+(n-k)b+2kc = (m-1)a+nb+k(2c-a-b)$. Now, under the condition $\nu^2 > 0$ and/or $\sigma_1^2 \neq \sigma_2^2$, we have $2c > a+b$. So, irrespective of whether $\mathcal{H}$ starts (and ends) with F or G, the cost of $\mathcal{H}$ is minimum when $k=1$. Therefore, $\mathcal{H}$ being the SHP, cannot have more that three runs, or in other words $P(T_{m,n} > 3 \mid T_{m,n} \text{ is odd }) \rightarrow 0$ as $d \rightarrow \infty$.

In case (ii), we have $T_{m,n} = 2k$ for some positive integer k. In this case, there are $m-k$ '**XX**-type' edges, $n-k$ '**YY**-type' edges and $2k-1$ '**XY**-type' edges in $\mathcal{H}$. So, the total cost of $\mathcal{H}$ converges to $(m-k)a+(n-k)b+(2k-1)c = (m-1)a+(n-1)b+2c+(k-1)(2c-a-b)$, which is minimum when $k=1$. Therefore, $P(T_{m,n} > 2 \mid T_{m,n} \text{ is even }) \rightarrow 0$ as $d \rightarrow \infty$. Now, combining case (i) and case (ii), we get $P(T_{m,n} > 3) \rightarrow 0$ as $d \rightarrow \infty$.

Note that $m!\,n!/(m+n-1)! < 0.05$ for all $m,n \geq 5$. So, for large dimensional consistency of the MR test with 5% nominal level, it is enough to have 5 observations from each distribution. However, under the condition of Theorem 1, if ν^2 exceeds $|\sigma_1^2 - \sigma_2^2|$, from the proof of the theorem, it is easy to check that (by comparing the cost of $\mathcal{H}$ in case (i) and case (ii)) $P_{H_1}(T_{m,n} = 2) \rightarrow 1$ as $d \rightarrow \infty$. So, in that case, it is enough to have $m,n \geq 4$ that ensures $P_{H_0}(T_{m,n} = 2) = 2/\binom{m+n}{m} < 0.05$. Box plots in Figure 3(b) show the distributions of $T_{m,n}$ for different choices of d in the location problem discussed in Section 3. This figure clearly shows that $T_{m,n}$ converges to 2 as d increases. But, if we have $\nu^2 < |\sigma_1^2 - \sigma_2^2|$, $T_{m,n} \xrightarrow{P} 3$ as $d \rightarrow \infty$. We observed it in the scale problem discussed in Section 3 (see Figure 3(d)). If $\sigma_1^2 > \sigma_2^2$ (or $\sigma_1^2 < \sigma_2^2$, respectively), the SHP starts and ends with observations from F (or G, respectively) with all observations from G (or F, respectively) in the middle.

Let us now investigate the high dimensional behavior of the FR test. Under (A1)-(A3), the performance of this test highly depends on the ordering of three types of distances ('**XX**', '**XY**' and '**YY**'). Note that if $\nu^2 > |\sigma_1^2 - \sigma_2^2|$ (i.e., $\sigma_1\sqrt{2}, \sigma_2\sqrt{2} < \sqrt{\sigma_1^2 + \sigma_2^2 + \nu^2}$), for large d, all '**XY** type' distances become larger than all '**XX** type' and all '**YY** type' distances. In that case, each and every observation from F (or G, respectively) tend to have all of its first $m-1$ (or $n-1$, respectively) nearest neighbors from F (or G, respectively) itself. As a result, the FR test statistic $T_{m,n}^{FR}$ attains its minimum value 2 with probability tending to one. We observed it in the location problem in Section 3 (see Figure 3(a)), where we had $\sigma_1^2 = \sigma_2^2 = 1$ and $\nu^2 = 0.09$. So, in this case, unless m and n are very small, the power of the FR test converges to 1 as d tends to infinity. However, the situation gets completely changed if $\nu^2 < |\sigma_1^2 - \sigma_2^2|$ (i.e., either $\sigma_1\sqrt{2}$ or $\sigma_2\sqrt{2}$ exceeds $\sqrt{\sigma_1^2 + \sigma_2^2 + \nu^2}$). Without loss of generality, let us assume $\sigma_1^2 - \sigma_2^2 > \nu^2$ as it was in the case of scale problem in Section 3. In this case, each observation from G has its first $n-1$ neighbors from G as before, but each observation from F has all of its first n nearest neighbors from G. As a

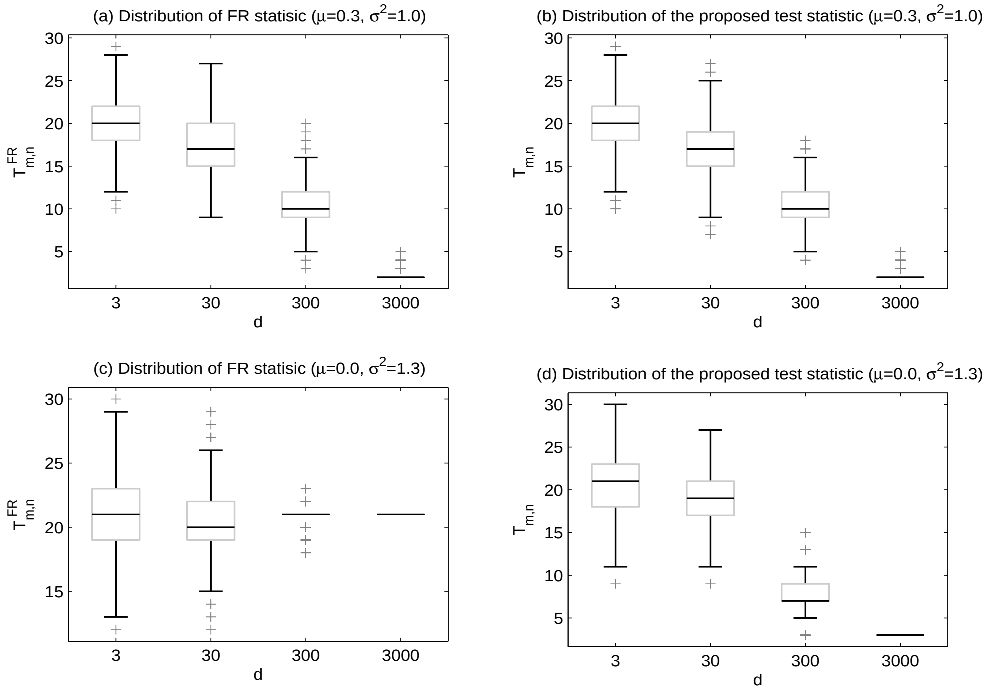

Figure 3: Distribution of FR and proposed test statistics for varying choices of d.

result, for higher values of d, $T^{FR}_{m,n}$ converges (in probability) to $m+1$ (see Figure 3(c)), which is equal to (even bigger than) its expected value under H_0 if $m=n$ ($m>n$), and much higher than the cut-off. This is precisely the reason why this test failed to yield satisfactory performance in the scale problem. In fact, in such cases, depending on m and n, the power of this test may even tend to zero as d tends to infinity. These results are summarized in the following theorem.

Theorem: *Suppose that F and G both satisfy the assumptions (A1)-(A3).*
(i) If $\nu^2 > |\sigma_1^2 - \sigma_2^2|$ and $\max\{\lfloor N/n \rfloor, \lfloor N/m \rfloor\}/\binom{m+n}{m} \leq \alpha$, the power of the FR test of level α converges to 1 as $d \to \infty$ (Here, $\lfloor r \rfloor$ denotes the highest integer $\leq r$).
(ii) If $\nu^2 < \sigma_1^2 - \sigma_2^2$ and $m/n > (1+\alpha)/(1-\alpha)$ (interchange F and G if required), the power of the FR test of level α converges to 0 as $d \to \infty$.

Proof: (i) From our previous discussion, it is clear that under the condition $\nu^2 > |\sigma_1^2 - \sigma_2^2|$, $T^{FR}_{m,n} \xrightarrow{P} 2$ as $d \to \infty$ (see Figure 3(a)). So, there is a subtree $\mathcal{T}_1$ on m vertices correspond to m observations from F and another subtree $\mathcal{T}_2$ on n vertices correspond to n observations from G. These two subtrees are connected by an edge $e = \{uv\}$, where u and v correspond to two vertices of $\mathcal{T}_1$ and $\mathcal{T}_2$, respectively (see Figure 4). Now, let us compute the probability $P(T^{FR}_{m,n} = 2)$ under the permutation distribution of $T^{FR}_{m,n}$. First note that if $\mathcal{T}_1$ and $\mathcal{T}_2$ both contain some vertices labeled as F and some labeled as G, $T^{FR}_{m,n}$ cannot be 2. So, if $m=n$, there are only two possibilities. Either all vertices of $\mathcal{T}_1$ or all vertices of $\mathcal{T}_2$ should be labeled as F (see Figure 4(a)). Therefore, in that case, $P(T^{FR}_{m,n} = 2)$ turns out to be $2/\binom{m+n}{m}$. Now, without loss of generality, let us assume $m > n$. First note that in this case, all vertices of $\mathcal{T}_2$ should have the same label. If all of them are labeled as G, all vertices of $\mathcal{T}_1$ will get label F (see Figure 4(b)). If all vertices of $\mathcal{T}_2$ are labeled as F, in order to count the number of favourable cases, first note that u must have the label F. Also, at most one of its neighbors (vertices that share an edge with u) can have label G. Suppose w ($w \neq u$) is the neighbor having the label G. Consider the collection C_w of all vertices in $\mathcal{T}_1$ that connect to u through w. Note that all vertices in this collection (that includes w itself) should have label G, and no other vertices in $\mathcal{T}_1$ can have the label G. So, the cardinality of C_w must be n. Similarly, the other neighbors of u can label G only if the corresponding collection has cardinality n. So, if all k neighbors (including v) of u has cardinality n, the vertex w can be chosen in $k-1$ different ways, and the total number of favourable cases turns out to be k (including the one, where all vertices of $\mathcal{T}_2$ has label G). If u does not have any neighbor labeled as G, instead of u, the same argument can be used on each of the neighbors of u barring v. Note that in order to have these k favorable cases, we need if $kn+1 \leq N$ or $(N-1)/n > k$. So, we cannot have more than $\lfloor (N-1)/n \rfloor$ favourable cases. Similarly, if $n > m$, the number of favourable cases cannot exceed $\lfloor (N-1)/m \rfloor$. Recall that if $N/n = N/m = 2$ (i.e., $m = n = 2$), the number of favourable cases is 2. So, combining all these results, we get an upper bound for $P(T^{FR}_{m,n} = 2)$, which is given by $k/\binom{N}{m}$, where $k = \max\{\lfloor N/n \rfloor, \lfloor N/m \rfloor\}$. If this upper bound is smaller than α, the power of the FR test of level α converges to 1 as $d \to \infty$.

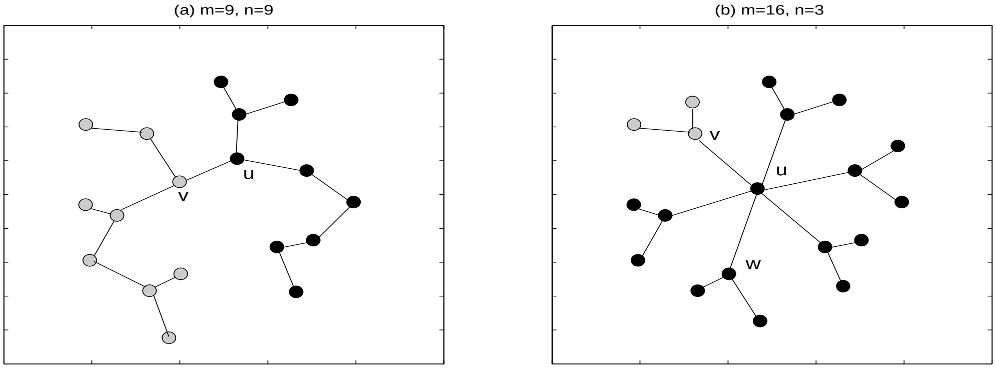

Figure 4: Minimal spanning trees with $T^{FR}_{m,n} = 2$.

(ii) It is easy to check that (also clear from our discussion) under the given condition $T = T^{FR}_{m,n} - 1$ converges to m in probability (see Figure 3(c)). Note that T is a non-negative random variable, and $E(T \mid \mathcal{Z})$, the conditional expectation of the permutation distribution of T given the data $\mathcal{Z}$=$\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m, \mathbf{y}_1, \mathbf{y}_2 \ldots, \mathbf{y}_n\}$ does not depend on $\mathcal{Z}$ (see e.g., Friedman and Rafsky), and we have $E(T \mid \mathcal{Z}) = \frac{2mn}{N}$ $\forall$ $\mathcal{Z}$, where $N = m+n$. Therefore, using the Markov inequality, we have $P(T \geq m \mid \mathcal{Z}) \leq 2n/N \Rightarrow P(T < m \mid \mathcal{Z}) \geq (m-n)/N$. Now, $m/n > (1+\alpha)/(1-\alpha)$ implies $(m-n)/N > \alpha$ and that completes the proof.

Here, one should note that part (ii) of Theorem 2 gives only a sufficient condition when the FR test fails. This test may fail in many other cases. For instance, in the scale problem in Section 2, we had $m = n = 20$ (i.e., $m/n = 1$), but the power of FR test dropped down to 0 as d increased. We will see a few more examples in the next section.

5 Results from the analysis of simulated data sets

In this section, we carry out simulation studies to compare the performance of the proposed test with some of the popular nonparametric two-sample tests available in the literature. Along with the FR test and the NBM test, here we consider two other tests for comparison, the test based on nearest neighbor (NN) type coincidences (see, e.g., Schilling, 1986; Henze, 1988) and the test proposed by Baringhaus and Franz (2004). In future, we will refer to them as the NN test and the BF test, respectively. In each of these simulated examples, we generated equal number of observations from the two distributions to constitute the sample. We carried out our analysis for two different choice of sample sizes, $m = n = 20$ and $m = n = 50$. As we have mentioned before, the MR test and the NBM test have the distribution-free property. For FR, NN and BF tests, in the case of $m = n = 20$, we used the conditional tests based on the permutation principle. In the case of $m = n = 50$, we used both, the conditional test and the test based on the large sample distribution of the test statistic. In each case, the best one (which happened to be the conditional test in almost all cases) has been reported here. The codes for the large sample tests based on NN and BF statistics are available at the R packages 'MTSKNN' and 'cramer', respectively. In all other cases, we used our own codes. Each experiment was repeated 500 times as before, and the estimated powers of different tests for various choices of d are reported in Table 1. Throughout this article, we assume that all tests have 5% nominal level.

We begin with the location and the scale problems considered in Section 3 (call them Example-1 and Example-2, respectively). In the preceding section, we observed that in both these cases, the proposed test had better performance than the FR test and the NBM test. This is what we observed also in Table 1, which shows the powers of these tests for three different values of d (30, 60 and 90). In the location problem, the BF test had the best performance followed by the NN test. The MR test had the third best performance in this example. But, in the case of scale problem, the MR test outperformed all other two-sample tests considered here. In view of the theoretical results discussed in Section 4, good performance of the MR test and poor performance of the FR test were quite expected in this example. One should also note that like the FR test, the power of the NN test also dropped down as the dimension increased. The reason behind such behavior of the NN test can be explained using the same type of argument as used in part (ii) of Theorem 2 (see also Biswas and Ghosh, 2013). The BF test and the NBM test had powers increasing with d, but the rate of increment was very low compared to that of the MR test. The reason behind the poor performance of the BF test in scale problems has been discussed in Biswas and Ghosh (2013).

Now, let us consider some examples, where at least one of the conditions of Theorem 1 fails to hold. This theorem gives us a fair idea about the high dimensional performance of the MR test under the assumptions (A1)-(A3). So, here we consider these examples to check how this test and other nonparametric tests perform when at least one of these conditions is violated. Example-3 and Example-4 deal with two multivariate normal distributions, where all component variables in F and G follow the standard normal distribution. So, in these two examples, F and G differ only in their

Table 1: Observed powers of two-sample tests (with 5% nominal level) in simulated data sets.

		$m = n = 20$					$m = n = 50$				
		FR	NBM	BF	NN	MR	FR	NBM	BF	NN	MR
Ex.-1	$d = 30$	0.292	0.268	**0.834**	0.492	0.346	0.618	0.502	**0.998**	0.842	0.684
	$d = 60$	0.484	0.386	**0.968**	0.742	0.542	0.880	0.782	**1.000**	0.978	0.914
	$d = 90$	0.616	0.518	**0.998**	0.872	0.664	0.962	0.886	**1.000**	0.994	0.980
Ex.-2	$d = 30$	0.036	0.092	0.098	0.070	**0.216**	0.058	0.114	0.142	0.078	**0.384**
	$d = 60$	0.024	0.120	0.122	0.052	**0.472**	0.028	0.130	0.278	0.068	**0.782**
	$d = 90$	0.008	0.138	0.150	0.042	**0.618**	0.016	0.162	0.416	0.040	**0.952**
Ex.-3	$d = 30$	0.388	0.322	0.080	**0.478**	0.436	0.940	0.874	0.152	**0.990**	0.978
	$d = 60$	0.420	0.384	0.076	**0.582**	0.486	0.958	0.910	0.136	**0.998**	0.988
	$d = 90$	0.444	0.392	0.050	**0.588**	0.494	0.972	0.932	0.130	**1.000**	0.992
Ex.-4	$d = 30$	0.072	0.122	0.060	0.102	**0.176**	0.128	0.178	0.092	0.228	**0.422**
	$d = 60$	0.052	0.128	0.054	0.106	**0.366**	0.100	0.230	0.086	0.196	**0.792**
	$d = 90$	0.032	0.148	0.066	0.094	**0.550**	0.096	0.272	0.094	0.202	**0.942**
Ex.-5	$d = 30$	0.000	0.306	0.492	0.010	**0.848**	0.000	0.710	0.922	0.012	**0.996**
	$d = 60$	0.000	0.352	0.598	0.000	**0.956**	0.000	0.800	0.974	0.000	**0.998**
	$d = 90$	0.000	0.434	0.674	0.000	**0.990**	0.000	0.852	0.990	0.000	**1.000**
Ex.-6	$d = 30$	0.088	0.142	0.038	0.178	**0.284**	0.176	0.422	0.060	0.350	**0.672**
	$d = 60$	0.084	0.264	0.024	0.206	**0.400**	0.172	0.684	0.062	0.462	**0.926**
	$d = 90$	0.088	0.390	0.020	0.252	**0.556**	0.194	0.880	0.056	0.536	**0.978**

correlation structures. In Example-3, F and G have the scatter matrices $\mathbf{\Sigma}_F = (((0.35)^{|i-j|}))_{d\times d}$ and $\mathbf{\Sigma}_G = (((-0.35)^{|i-j|}))_{d\times d}$, respectively. In Example-4, while all off-diagonal elements of $\mathbf{\Sigma}_F$ are 0.1, those of $\mathbf{\Sigma}_G$ are 0.3. Note that in Example-3, (A1)-(A3) hold both for F and G, but we have $\nu^2 = 0$ and $\sigma_1^2 = \sigma_2^2$. In Example-4, again we have $\nu^2 = 0$ and $\sigma_1^2 = \sigma_2^2$, and the ρ-mixing condition in (A2) is also violated. In Example-3, the NN test had the best performance followed by the MR test. The FR test and the NBM test had slightly low powers. In this example, the BF test failed to compete with other methods. While the powers of other tests increased with the dimension that of the BF test dropped down. In Example-4, the MR test clearly outperformed all its competitors. The NBM test had the next best performance, but even its power was not at all comparable to that of the MR test.

In Example-5, the two distributions F (multivariate normal with the location parameter $\mathbf{0}$ and the scatter matrix $\mathbf{I}_d$) and G (multivariate t-distribution with 3 degrees of freedom, the location parameter $\mathbf{0}$ and the scatter matrix $\frac{1}{3}\mathbf{I}_d$) have the same mean vector and the same dispersion matrix, but they differ in their shapes. In this example, again we have $\nu^2 = \sigma_1^2 - \sigma_2^2 = 0$, and the assumptions (A1) and (A2) are violated in G. The MR test had excellent performance in this example as well. While the FR test and the NN test both failed to reject the null hypothesis even in a single occasion, the MR test rejected H_0 in more than 99.5% of the cases. The BF test had somewhat comparable performance in the case of $m = n = 50$, but in the case of $m = n = 20$, its power was much lower than that of the MR test.

Next, we consider an example (Example-6), where F is an equal mixture of two normal distributions $N_d(0.3\ \mathbf{1}_d,\ \mathbf{I}_d)$ and $N_d(-0.3\ \mathbf{1}_d,\ 4\mathbf{I}_d)$, and G is also an equal mixture of two normal distributions $N_d(0.3\ \boldsymbol{\beta}_d,\ \mathbf{I}_d)$ and $N_d(-0.3\ \boldsymbol{\beta}_d,\ 4\mathbf{I}_d)$. Here $\mathbf{1}_d = (1, 1, \ldots, 1)^{'}$ denotes the d-dimensional vector with all elements unity, and $\boldsymbol{\beta}_d = (1, 1, \ldots, 1-1, -1, \ldots, -1)^{'}$ has first $d/2$ elements equal to 1 and the rest are -1. One can check that the assumption (A2) is violated in both of these distributions. In this example, the MR test again outperformed its competitors. The BF test, which is based on average of intra-class and inter-class distances, had very poor performance. Among the rest, the NBM test had the next highest power in most of the cases. Instead of normal mixtures, we carried out this experiment also with mixtures of t-distributions and mixtures of Cauchy distributions, but the basic finding was almost the same.

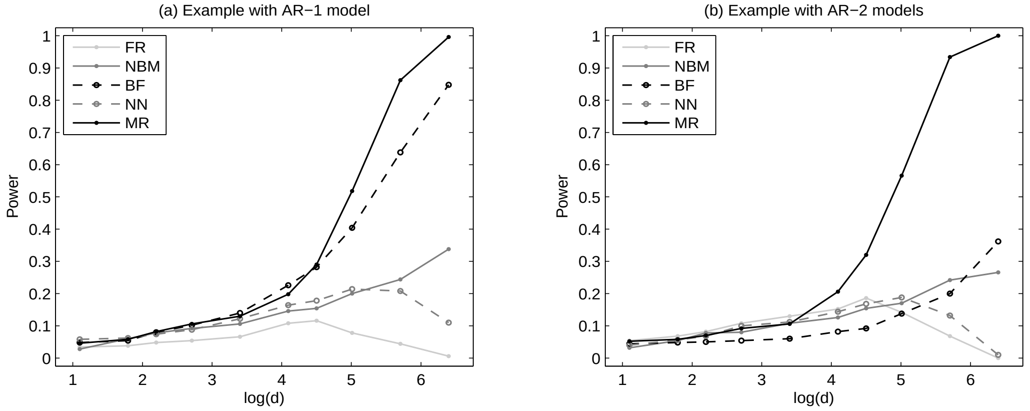

Figure 5: Powers of FR, NBM, BF, NN and MR tests for varying choices of d.

Finally, we consider two examples with auto-regressive (AR) processes. In the first example, the observations in F were generated using the AR(1) model $X^{(t)} = 0.25 + 0.3X^{(t-1)} + U_t$ for $t = 1, 2, \ldots, 500$, where $X^{(0)}, U_1, U_2, \ldots, U_{500} \overset{i.i.d.}{\sim} N(0, 1)$, and the observations in G were generated using another AR(1) model $Y^{(t)} = 0.25 + 0.5Y^{(t-1)} + V_t$, where $Y^{(0)}$, V_1, $V_2, \ldots, V_{500} \overset{i.i.d.}{\sim} N(0, 1)$. In the other example, the observations in F were generated using the AR(2) model $X^{(t)} = 0.3X^{(t-1)} + 0.2X^{(t-2)} + U_t$ for $t = 1, 2, \ldots, 500$, and those in G were generated using the model $Y^{(t)} = 0.4Y^{(t-1)} + 0.3Y^{(t-2)} + V_t$ for $t = 1, 2, \ldots, 500$, where $X^{(0)}, X^{(-1)}, Y^{(0)}, Y^{(-1)}, U_1, U_2, \ldots, U_{500}, V_1, V_2, \ldots, V_{500}$ are all i.i.d. standard normal variates. Note that the conditions (A1)-(A3) hold in these examples. In the example with AR(1) processes, F and G differ both in their locations and scales, but in the example with AR(2) processes, they differ only in their scales. In both these examples, we generated 20 observations from each class to form the sample, and the experiment was repeated 500 times as before. We performed this experiment for various choices of d starting from 3 to 3000, and the results are presented in Figure 5. The superiority of the MR test for high dimensional data is quite transparent from this figure, especially in the example with AR(2) processes. We carried out these experiments with AR(1) and AR(2) processes for $m = n = 50$, but the superiority of the MR test was evident even in those cases.

6 Results from the analyses of benchmark datasets

In this section, we analyze five real benchmark data sets for further assessment of the proposed method. The Trace data set is obtained from the UCR time series classification/clustering page (http://www.cs.ucr.edu/ ∼eamonn/time_series_data/). The Colon data set is available at the R package 'dprep'. The rest of the data sets are taken from the UCI machine learning repository (http://archive.ics.uci.edu/ml/datasets/). Detailed descriptions of these data sets are available at these repositories. The Trace data set contains observations from four different classes. Here we consider two testing problems related to this data set, one between the first and the second classes, and the other between the third and the fourth classes. Henceforth, we will refer to these two subsets of Trace data as Trace data-1 and Trace data-2, respectively. All these benchmark data sets have been extensively used in the literature, mainly in context of supervised classification. It is also well known that in all these cases, we have reasonable separability between two competing classes. So, here we can assume the alternative hypothesis $H_A : F \neq G$ to be the true, and different tests can be compared on the basis of their power functions. However, if we use the whole data set for testing, any test will either reject H_0 or accept it. In fact, because of the reasonable separability between the two classes, most of the tests are expected to reject H_0. So, based on that single experiment, it is difficult to compare among different test procedures. Therefore, in each of these cases, we repeated the experiment 500 times based on 500 different subsets chosen from the whole data set at random, and the estimated power of different tests are computed based on these 500 subsets. In all these cases, we chose the equal number of observations from the two classes to form these subsets, and the subset sizes are kept to be small compared to the dimension of the data to make the testing problem more challenging. Throughout this section, for FR, BF and NN tests, we used the conditional test based on 500 permutations. For each of these data sets, we repeated the experiment with different subset sizes, and the powers of different tests are shown in Figure 6.

Ionosphere data set contains 34-dimensional observations from two classes, namely, 'Good' and 'Bad', which correspond to 'Good' radar return and 'Bad' radar return, respectively. Radar data were collected by a system in Goose Bay that consists of a phased array of 16 high-frequency antennas. The targets were free electrons in the ionosphere. Radar returns showing evidence of some type of structure in the ionosphere are terms as 'Good', and the returns those do not show any evidence are termed as 'Bad'. There were 17 pulse numbers for the Goose Bay system, and for each pulse number, there were two attributes. There were 126 instances of 'Good' and 225 instances of 'Bad' radar returns (see Sigillito et. al. (1989) or the UCI machine learning repository for details). In this data set, the MR test and the BF test had better performance than their competitors (see Figure 6(a)). For sample sizes less than 20, the BF test had a slight edge over the MR test, but the MR test had higher power afterwards. These two tests had power 1 for samples of size 40 or higher. The performances of other three tests were also comparable. Among them, the NN test had higher power compared to NBM and FR tests.

Next, we analyze the Sonar data set. This data set was used by Gorman and Sejnowski (1988) in their study of classification of sonar signals using a neural network. It contains 111 patterns obtained by bouncing sonar signals off a 'metal cylinder' and 97 patterns obtained from 'rocks' at different angles and under different conditions. The transmitted sonar signal is a frequency-modulated chirp, rising in frequency. Each pattern is a set of 60 numbers in the range 0.0 to

1.0, where each number represents the energy within a particular frequency band, integrated over a certain period of time. In this data set, the NN test had the best overall performance closely followed by the MR test (see Figure 6(b)). In all cases, the difference between their powers was less than 0.02. The FR test also had comparable performance. The BF test had the highest power for samples of size 10, but it was outperformed by the NN and the MR tests for all larger sample sizes. The NBM test did not have satisfactory performance in this data sets. For instance, when all other tests reached the maximum power 1, it had power less 0.3.

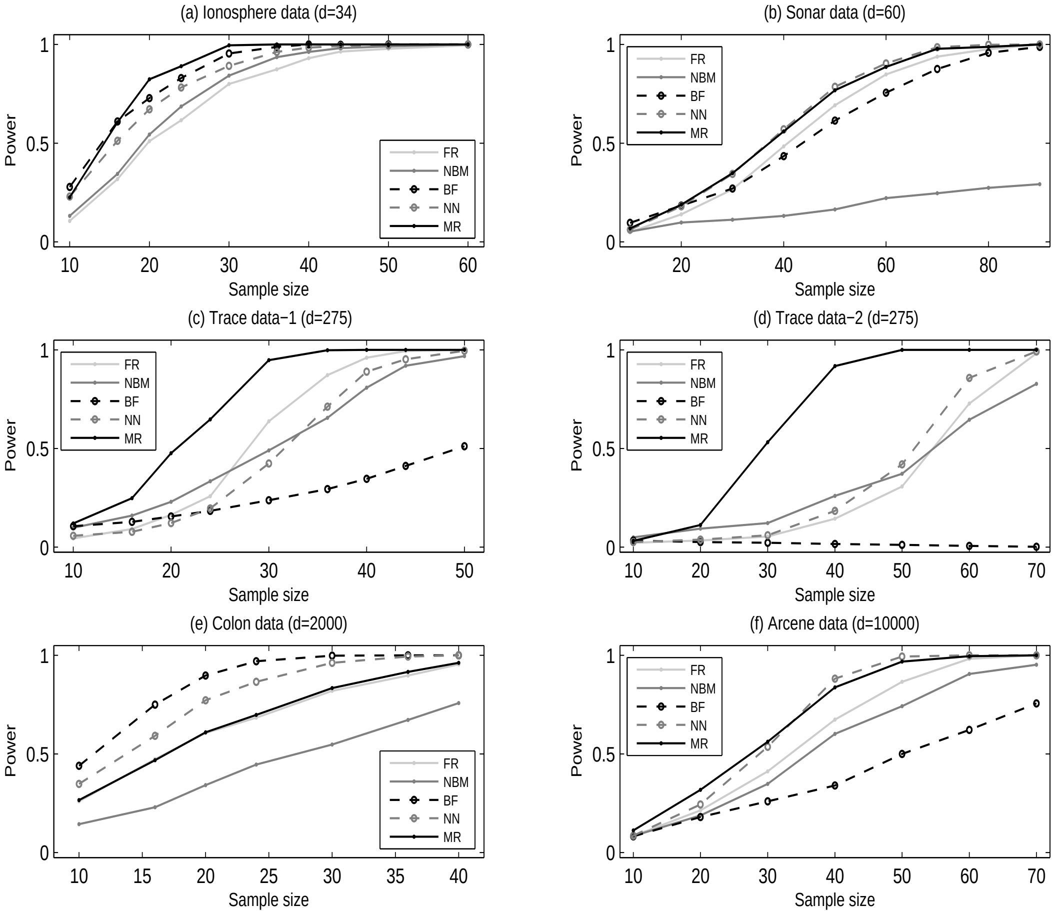

Figure 6: Powers of different tests in benchmark data sets.

The Trace data set was designed to simulate instrumentation failures in a nuclear power plant. It is a subset of the transient classification benchmark of trace project. The full data set consists of 16 classes each containing 50 instances, where each instance has four features. The Trace data set (subset) only uses the second feature of class 2 and 6, and the third feature of class 3 and 7, which are the first, the second, the third and the fourth class of this data set, respectively. This data set contains 200 instances, 50 for each class. All instances are linearly interpolated to have the same length of 275 data points (see the UCR time series classification/clustering page for details). We considered all $\binom{4}{2}$ pairs of classes separately for testing, but in four out of these six cases, because of high separability between the two classes, almost all tests attained power 1 even when very small samples were used. So, here we report the results only for two testing problems, one between the first and the second classes (referred to as Trace data-1), and the other between the third and the fourth classes (referred to as Trace data-2). In both these cases, our proposed test had excellent performance, and it led to substantially higher power than all other tests considered here (see Figures 6(c) and 6(d)). The BF test had very poor performance in these data sets, especially in Trace data-2, where its power had an unnatural decreasing trend with increasing sample size. Note that in Trace data-2, twice the average interclass-distance was smaller than the sum of the averages of intra-class distances for the two classes. This was the reason for this surprising behavior of the power function.

Next, we consider two high-dimensional data sets, where the data dimensions are larger than 1000. The Colon data set is a microarray gene expression data set that contains expression levels for 2000 genes for each of 62 samples, 40 from colon cancer tissues and 22 from normal tissues. The task was to distinguish between these two types of tissues based on their gene expression levels. Detailed description of this data set can be found in Alon et. al. (1999), and the data can be freely downloaded from http://microarray.princeton.edu/oncology/affydata/index.html. In this data set, the BF test yielded the best performance, while the NN test had the second position (see Figure 6(e)). Two multivariate run tests, the FR test and the MR test had almost similar performance, and they outperformed the NBM test. Finally, we analyze the Arcene data set. This is one of five data sets of the NIPS 2003 feature selection challenge. It was obtained by merging three mass-spectrometry data sets. All the data consist of mass-spectra obtained with the SELDI technique. The samples include cancer patients (ovarian or prostate cancer) and healthy patients. There were 7000 original features indicating the abundance of proteins in human sera having a given mass value. In addition to that, 3000 distractor features with no predictive power or 'probes' were added to increase the number of features to 10000. More details on this data set can be obtained in Guyon et. al. (2005). There were separate training, test and validation sets in the UCI machine learning repository. For our analysis, we use random subsets from the training set consisting of 44 cancer patients and 56 healthy patients. In this data set, the MR test and the NN test outperformed all other tests considered here. The MR test had an edge over the NN test for small sample sizes, but for samples of size 40 and 50, the NN test had the highest power. Both of them had power 1 for samples of size 60 or higher. If not better, the overall performance of the MR test in these benchmark data sets was comparable to the popular two-sample tests available in the literature. The analysis of these high dimensional data sets clearly indicates that the MR test can be conveniently used for the analysis of high dimension low sample size data.

7 Concluding remarks

In this article, we have proposed a multivariate generalization of the univariate Wlad Wolfowitz run test. Unlike Freidman and Rafsky's (1979) multivariate run test, this proposed test has the distribution-free property, and more interestingly, the distribution of the proposed test statistic $T_{m,n}$ exactly matches with that of the univariate run statistic. So, in some sense, it can viewed as the most natural generalization of the univariate run test. This test is based on pairwise distances between the sample observations, and the test statistic is invariant under location change, rotation and homogeneous scale transformation. This test can be conveniently used for high dimensional data even when the data dimension exceeds the sample. Our theoretical studies in Section 4 establish good power properties of the proposed test for such high dimensional data. Using our analysis of simulated and real benchmark data sets, we have also amply demonstrated the usefulness of the proposed test in high dimension low sample size situations.

This proposed idea based on SHP can be used as a general recipe for distribution-free multivariate generalizations of many other univariate rank based tests. For instance, using this idea, one can develop distribution-free multivariate versions of the Wilcoxon-Mann-Whitney statistic or the Kolmogorov-Smirnov statistic, which can be used even when the dimension of the data is larger than the sample size. The first one is useful for multivariate two-sample location problem, where one can show that unless the sample size is too small (i.e., if $2/\binom{m+n}{m}$ is not smaller than α), under the regularity conditions (A1)-(A3), the power of the proposed test converges to unity as the dimension increases. The latter can be used for general two-sample problem, and similar high dimensional consistency results can be proved for that test as well. However, one needs to investigate the empirical performance of these tests in high dimensional data.

Acknowledgement

The authors would like to thank Dr. Sasthi C. Ghosh for his valuable suggestions and also for helping them in writing the computer programs.

References

[1] Ahn, J., Marron J. S., Muller K. M., and Chi Y.-Y. (2007). The high-dimension, low sample-size geometric representation holds under mild conditions. *Biometrika*, **94**, 760-766.

[2] Alon U., Barkai, M., Notterman, D. A., Gish, K., Ybarra, S., Mack, D. and Levine, A. J. (1999) Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Nat. Acad. Sci., USA*, **96**, 6745–6750.

[3] Andrews, D. W. K. (1988). Laws of large numbers for dependent non-identically distributed random variables. *Econometric Theory*, **4**, 458–467.

[4] Baringhaus, L. and Franz, C. (2004). On a new multivariate two-sample test. *J. Multivariate Anal.*, **88**, 190-206.

[5] Baringhaus, L. and Franz, C. (2010). Rigid motion invariant two-sample tests. *Statist. Sinica*, **20**, 1333-1361.

[6] Bickel, P. J. (1969) A distribution free version of the Smirnov two sample test in the p-variate case. *Ann. Math. Statist.*, **40**, 1–23.

[7] Biswas, M. and Ghosh, A. K. (2013) A nonparametric two-sample test applicable to high dimensional data. Submitted for publication.

[8] Choi, K. and Marden, J. (1997). An approach to multivariate rank tests in multivariate analysis of variance. *J. Amer. Statist. Assoc.*, **92**, 1581-1590.

[9] de Jong, R. M. (1995) Laws of large numbers for dependent heterogeneous processes. *Econ. Theory*, **11**, 347 – 358.

[10] Ferger, D. (2000) Optimal tests for the general two-sample problem. *J. Multivariate Anal.*, **74**, 1-35.

[11] Friedman, J. H. and Rafsky, L. C. (1979). Multivariate generalizations of the Wald-Wolfowitz and Smirnov two-sample tests. *Ann. Statist.*, **7**, 697-717.

[12] Garey, M. and Johnson, D. (1979) *Computers and Intractability: A Guide to the Theory of NP Completeness.* W.H. Freeman and Co., San Francisco.

[13] Gibbons, J. D. and Chakraborti, S. (2003). *Nonparametric Statistical Inference.* Marcel Dekker, New York.

[14] Gorman, R. P. and Sejnowski, T. J. (1988). Analysis of hidden units in a layered network trained to classify sonar targets. *Neural Networks*, **1**, 75–89.

[15] Guyon, I.M., Gunn, S.R., Ben-Hur, A. and Dror, G. (2005). Result analysis of the NIPS 2003 Feature Selection Challenge. *Adv. Neural Info. Proc. Sys.*, **17**, 545-552.

[16] Hall, P., Marron, J. S., and Neeman, A. (2005). Geometric representation of high dimension, low sample size data. *J. Royal Statist. Soc. Ser. B*, **67**, 427-444.

[17] Hall, P. and Tajvidi, N. (2002). Permutation tests for equality of distributions in high dimensional settings. *Biometrika*, **89**, 359-374.

[18] Henze, N. (1988). A multivariate two-sample test based on the number of nearest neighbor type coincidences. *Ann. Statist.*, **16**, 772-783.

[19] Hettmansperger, T. P., Möttönen, J., and Oja, H. (1998). Affine invariant multivariate rank tests for several samples. *Statist. Sinica*, **8**, 785-800.

[20] Hettmansperger, T. P. and Oja, H. (1994). Affine invariant multivariate multi-sample sign tests. *J. Royal Statist. Soc. Ser. B*, **56**, 235-249.

[21] Jung, S. and Marron, J. S. (2009). PCA consistency in high dimension, low sample size context. *Ann. Statist.*, **37**, 4104-4130.

[22] Kruskal, J. B. (1956) On the shortest spanning subtree of a graph and the travelling salesman problem. *Proc. Amer. Math. Soc.*, **7**, 48–50.

[23] Lawler, E. L., Lenstra, J. K., Kan, A. H, G, R and Shmoys, D. B. (1985) The Travelling Salesman Problem. Wiley, New York.

[24] Liu, R. Y. and Singh, K. (1993). A quality index based on data depth and multivariate rank tests. *J. Amer. Statist. Assoc.*, **88**, 252-260.

[25] Liu, Z. and Modarres, R. (2011). A triangle test for equality of distribution functions in high dimensions. *J. Nonparametr. Statist.*, **23**, 605-615.

[26] Lu, B., Greevy, R., Xu, X., and Beck, C. (2011). Optimal non-bipartite matching and its statistical applications. *Amer. Statist.*, **65**, 21-30.

[27] Mardia, K. V. (1967) A nonparametric test for bivariate location problem. *J. Royal Statist. Soc., Ser. B*, **29**, 320-342.

[28] Möttönen, J. and Oja, H. (1995). Multivariate spatial sign and rank methods. *J. Nonparametr. Statist.*, **5**, 201-213.

[29] Oja, H. (2010). *Multivariate Nonparametric Methods with R: An Approach Based on Spatial Signs and Ranks.* Springer, New York.

[30] Oja, H. and Randles, R. H. (2004). Multivariate nonparametric tests. *Statist. Sci.*, **19**, 598-605.

[31] Puri, M. L. and Sen, P. K. (1971). *Nonparametric Methods in Multivariate Analysis.* Wiley, New York.

[32] Randles, R. H. and Peters, D. (1990). Multivariate rank tests for the two-sample location problem. *Comm. Statist. Theory Methods*, **19**, 4225-4238.

[33] Rosenbaum, P. R. (2005). An exact distribution-free test comparing two multivariate distributions based on adjacency. *J. Royal Statist. Soc. Ser. B*, **67**, 515-530.

[34] Rousson, V. (2002). On distribution-free tests for the multivariate two-sample location-scale model. *J. Multivariate Anal.*, **80**, 43-57.

[35] Schilling, M. F. (1986). Multivariate two-sample tests based on nearest neighbors. *J. Amer. Statist. Assoc.*, **81**, 799-806.

[36] Sigillito, V. G., Wing, S. P., Hutton, L. V. and Baker, K. B. (1989). Classification of radar returns from the ionosphere using neural networks. *Johns Hopkins APL Technical Digest*, **10**, 262–266.

[37] Wald, A. and Wolfowitz, J. (1940) On a test whether two samples are from the same distribution. *Ann. Statist.*, **11**, 147-162.

[38] Zech, G. and Aslan, B. (2003). A multivariate two-sample test based on the concept of minimum energy. *PHYSTAT*, 97-100.